

Prediction of Future Hospital Admissions – What is the Tradeoff Between Specificity and Accuracy?

Ieva Vasiljeva and Ognjen Arandjelović[†]

School of Computer Science
University of St Andrews
North Hough
St Andrews KY16 9SX
Scotland, Fife
United Kingdom

[†]ognjen.arandjelovic@gmail.com

Abstract

Large amounts of electronic medical records collected by hospitals across the developed world offer unprecedented possibilities for knowledge discovery using computer based data mining and machine learning. Notwithstanding significant research efforts, the use of this data in the prediction of disease development has largely been disappointing. In this paper we examine in detail a recently proposed method which has in preliminary experiments demonstrated highly promising results on real-world data. We scrutinize the authors' claims that the proposed model is scalable and investigate whether the tradeoff between prediction specificity (i.e. the ability of the model to predict a wide number of different ailments) and accuracy (i.e. the ability of the model to make the correct prediction) is practically viable. Our experiments conducted on a data corpus of nearly 3,000,000 admissions support the authors' expectations and demonstrate that the high prediction accuracy is maintained well even when the number of admission types explicitly included in the model is increased to account for 98% of all admissions in the corpus. Thus several promising directions for future work are highlighted.

1 Introduction

The concept of prognosis is of pervasive importance in medicine. As the very etymology of the word suggests (Greek *prognōsis*, *pro*-‘before’ + *gignōskein* ‘know’), prognosis concerns the *prediction* of medical outcomes. Examples include disease progression patterns, different aspects of life quality, mortality chances, and many others. Already recognized in the antiquity (amongst others by Ancient Egyptians, Sumerians, and Greeks)

the paradigm of prognosis – that is to say the underlying methodology – has changed dramatically. Initially predicated on highly limited and bias prone experience of practising physicians, the last couple of centuries have witnessed the development of a rigorous framework for collecting, interpreting, and using evidence – this is now widely referred to as evidence based medicine (or better yet, science based medicine [14]).

Recent advances in computing are promising to effect a major change in prognostic practice. In particular, the ability to record, store, and send across large distances massive amounts of patient data, together with major breakthroughs in machine learning and data mining, open the possibility of using artificial intelligence to analyse corpora far larger than ever before. The particular focus of the present paper is on the use of electronic medical records for the prediction of future complications likely to be experienced by a specific patient. Such predictions can aid in the understanding of disease aetiology, patient incentivization, and the allocation of hospital resources [3].

Considering the impact that accurate prediction of this type would have on finances, population health at large, and individual patients' well-being, it is unsurprising that the challenge of developing suitable models has already attracted significant research efforts. Broadly categorized, these fall under two umbrellas: (i) methods which involve explicit physiological modelling relevant to a specific condition [23, 11, 25], and (ii) methods which adopt a holistic strategy and approach the problem from a data-driven perspective, modelling the overall state of a patient's health [9, 16, 3]. An obvious limitation of the former group of methods is that by their very nature they are not readily generalizable. Moreover, it is likely that a restricted view of a single condition in isolation

is inadequate to model many modern diseases which have numerous comorbidities affecting a broad range of bodily systems [17]. To give but one example, some of the more common comorbidities of diabetes mellitus include macro-vascular ailments such as coronary artery disease, myocardial infarction, stroke, congestive heart failure, and peripheral vascular disease, micro-vascular complications such as retinopathy, nephropathy, and neuropathy, metabolic disorders such as dyslipidemia, non-alcoholic fatty liver disease, and obesity etc. Thus, holistic modelling approaches appear more attractive in principle. However, the task of developing models which are flexible enough to provide practically sufficient specificity, yet constrained enough to be learnable from real-world data, is a major challenge. Indeed, until recently the performance of different methods described in the literature has been disappointing. The present paper is motivated by a recently proposed method which has demonstrated impressive and at first sight highly promising results in preliminary experiments [3].

The key ideas underlying the adopted method and a sketch of its main technical aspects are described in the next section. Our main aim here was to scrutinize the original authors' expectation that the method would scale well i.e. that the predictive performance of the method, reported with explicit modelling of the 30 most frequent admission types only, could be maintained as a greater number of admission types is included in the model as most practical applications would demand. The original paper did not investigate this; rather, the number of salient, explicitly modelled admission types was set in an *ad hoc* manner to 30, explaining approximately 75% of the data corpus [3] (also see [5]). If our expectation of performance deterioration with an increased number of explicitly modelled admission types is correct, and if the rate of deterioration is high, the model could end up being of little practical significance: on the one end of the parameter spectrum the model would provide high accuracy but insufficient specificity for its predictions to be practically useful, and on the other high specificity but poor accuracy for its predictions to be relied upon. Thus the present analysis is necessary before any practical use can be considered.

2 Methods

In the present paper we conduct our investigation using the method recently proposed in [4, 3]. Our choice was motivated by its high predictive accuracy demonstrated on a large real-world data corpus. For the sake of completeness the key ideas and the main technical elements underlying the adopted method are summarized next; for comprehensive detail and a re-

lated discussion the reader is referred to the original publication.

2.1 Adopted model overview

Given a patient's hospital admission history H which comprises a sequence of admissions a_i :

$$H = a_1 \rightarrow a_2 \rightarrow \dots \rightarrow a_n, \quad (1)$$

where each a_i is a discrete variable whose value is a code from the International Statistical Classification of Diseases and Related Health Problems (ICD) [24], the adopted model predicts the most likely future admission a_{n+1}^* as:

$$a_{n+1}^* = \arg \max_{a \in A_{\text{ICD}}} p(H \rightarrow a), \quad (2)$$

where A_{ICD} is the set of all possible ICD codes. To make the estimation of the probability $p(H \rightarrow a)$ tractable, a patient's medical history H is represented using a fixed length binary vector $v(H)$. This representation bears some resemblance to the bag of words representation frequently used in text analysis [6, 7] and which has since been successfully adapted to various other application domains too [2, 1, 19, 21]. Each element in $v(H)$ encodes the presence (value 1) or lack thereof (value 0) of a specific salient admission (i.e. ICD code) in H , save for the last element which captures jointly all non-salient admission types. As in [3] saliency is determined by the frequency of the corresponding admission in the entire data corpus (n.b. different saliency criteria can be readily used instead, see Section 4). The probability $p(H \rightarrow a)$ in (2) is then estimated by superimposing a Markovian model [22, 15] on the space of history vectors which leads to $H \rightarrow a$ being interpreted as a transition from the state represented by $v(H)$ to the state represented by $v(H \rightarrow a)$. As usual the probabilities parameterizing the Markov model are learnt from a training data corpus. A conceptual illustration of the method is shown in Figure 1.

The key idea behind the described model is that it is the *presence* of past complications which most strongly predicts future ailments [18, 13, 12, 10], which allows for the space of states over which learning is performed to be reduced dramatically; in particular, this is achieved by employing a fixed length state representation and through binarization of its elements.

2.2 Data and experimental protocol

To make our results and conclusions directly comparable to those reported in the original paper which introduced the adopted history vector based method, we used the same large corpus of electronic medical

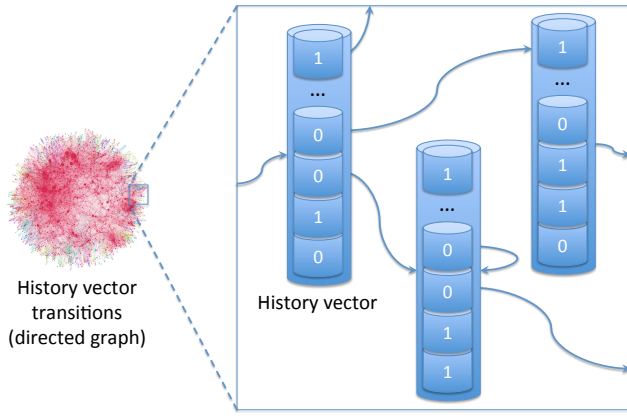


Figure 1: A conceptual summary of the adopted model which superimposes a Markovian model over a space of history vectors used to represent the medical state of a patient.

records collected by a local hospital. The data set contains entries corresponding to nearly 3,000,000 hospital admissions of 40,000 patients.

Our primary goal here is to examine how the predictive performance of the history vector based model is affected by the choice of the number of salient admission types (see Section 2.1). As in [3] we too assess the quality of a specific prediction by considering the rank of the ground truth admission type in the probability ordered list of predictions. Formally, let a_t be the ground truth admission type which follows a particular history H . Then the rank r of the ground truth admission type a_t is given by number of admission types which the model predicts as following H with at least the probability $p(H \rightarrow a_t)$:

$$r = |\{a : a \in A_{ICD} \wedge p(H \rightarrow a) \geq p(H \rightarrow a_t)\}|. \quad (3)$$

We used the same granularity of codes the original work described in [4, 3].

Furthermore, we adopt the usual ‘leave one out’ evaluation protocol whereby the performance of the method is tested with each patient’s data in turn and the model trained using the data of all other patients. To quantify the aggregate performance of the model for specific model parameter values (i.e. the number of salient admission types included in the history vector representation) we use two well known measures. These are the average rank (a special case of the average normalized rank [20] when the set of target matches is exactly equal to 1) and the normalized area under the cumulative match characteristic (CMC) curve. For each possible rank r ($r = 1 \dots n$, where n is the worst possible rank, equal to the number of admission types),

the CMC takes on the value equal to the proportion of predictions which predict the correct admission type at worst with the rank r [8]. The ideal performance results in the CMC having the value 1 across all ranks i.e. in each individual case the correct admission is ranked 1. The area under the curve is normalized so that it is equal to 1 in this ideal case.

3 Results and Discussion

We started by looking at the effect that changing the number of salient admission types, i.e. DRG codes with the corresponding (1-to-1) elements in the history vector, has on the area under the CMC curve. Our experimental results are captured by the plot in Figure 2(a). The plot can be readily seen to support the hypothesis put forward in Section 1 that predicted a decay in the adopted model’s prediction performance for an increasing number of explicitly modelled admission types. Notwithstanding this unwelcome qualitative observation, the major result is of a quantitative nature – the rate of the aforementioned decay is very slow indeed. Like many other natural phenomena the decay exhibits a power-law form with the associated exponent value which differs from 1 only 5 parts in 100,000 i.e. it is equal to $1 - 0.5 \times 10^{-5}$. The practical significance of this finding is better appreciated by considering the plot in Figure 2(b). This plot shows the variation in the area under the CMC curve as a function of the coverage of the entire admissions data corpus by the salient DRG codes. The outstanding performance of the adopted method is illustrated well by noting, for example, that the dimensionality of history vectors can be increased to explicitly model the number of most frequent DRG codes which cover over 91% of the data, with the predictive performance of the method dropping by a mere 0.5% as compared to the coverage of only 61%. Even 98% of data coverage results in a change of only 0.8%. Recall that in the original paper the authors used 30 DRG codes which accounted for 75% of the admissions in the corpus. Our results demonstrate that this was an overly conservative value.

We next examined the average prediction rank of the correct admission type, which offers further insight into the performance of the adopted method. As expected from the previous set of findings, the results summarized by the plots in Figs 3(a) and 3(b) corroborate the observation that an increase in the dimensionality of history vectors, a key parameter of the method, worsens performance. In this experiment this worsening is exhibited as an increase in the average rank (i.e. a greater number of incorrect predictions are made with a higher probability than the actual ground truth admission type). It is interesting to note the significance

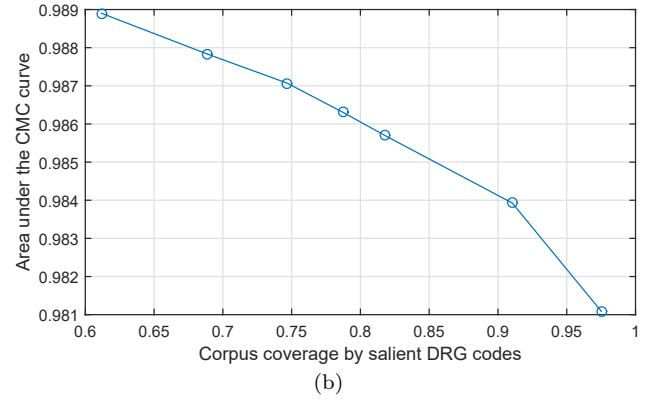
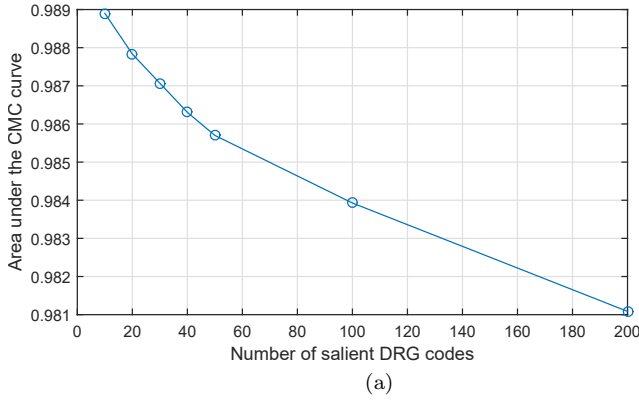


Figure 2: The normalized area under the cumulative match characteristic (CMC) curve.

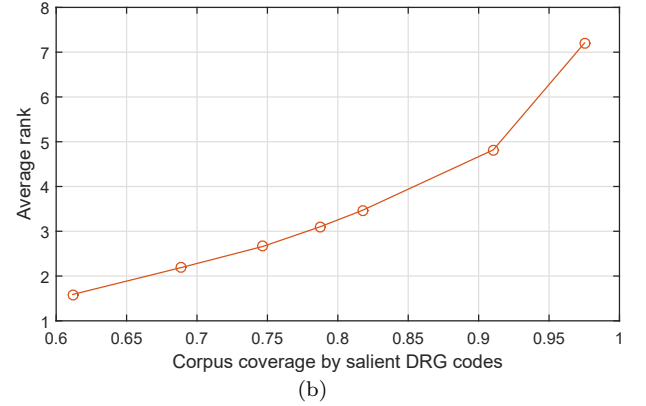


Figure 3: The average prediction rank of the correct admission type.

of what appears to be a much more rapid performance deterioration in terms of this performance measure in comparison with the area under the CMC curve discussed previously. For example, while the use of 200 vs. 10 most frequent DRG codes effects a reduction of only 0.5% in the area under the CMC curve, the corresponding change in the average rank of the correct admission type increases fivefold (from approximately 1.5 for 10 salient DRG codes, to approximately 7.3 for 200 salient codes). The explanation for this apparent discrepancy is in fact reassuring as it demonstrates that the most dramatic changes in the predicted rank happen for predictions which are already not very good i.e. the small number of bad predictions become even worse, rather than good predictions becoming bad.

Lastly, to examine in additional detail how an increase in the number of explicitly modelled admission type affects predictions, we looked at prediction rank histograms for different DRG codes and the corresponding changes as their number was changed. Figure 4(a) and 4(b) contrast the histograms for 20 and 50 salient admission types. It is remarkable to

observe that in both cases the histograms are virtually identical across different codes within the same model. Rather than being effected by sub-par histograms of the added DRG codes, the (small, as demonstrated previously) deterioration in predictive performance as the number of salient admission types is increased, is effected by slightly worse predictive performance uniformly distributed across different DRG codes. This is highly preferable in practice as it implies that for a fixed model complexity predictive power remains the same regardless of the patient’s ailment. Were it otherwise, the predictions would be more difficult to interpret and the model complexity more challenging to set appropriately as the model’s predictive performance would exhibit dependence on the nature of the health problems affecting a specific patient.

4 Summary and Conclusions

In this paper we considered a recently proposed computational model which uses electronic medical

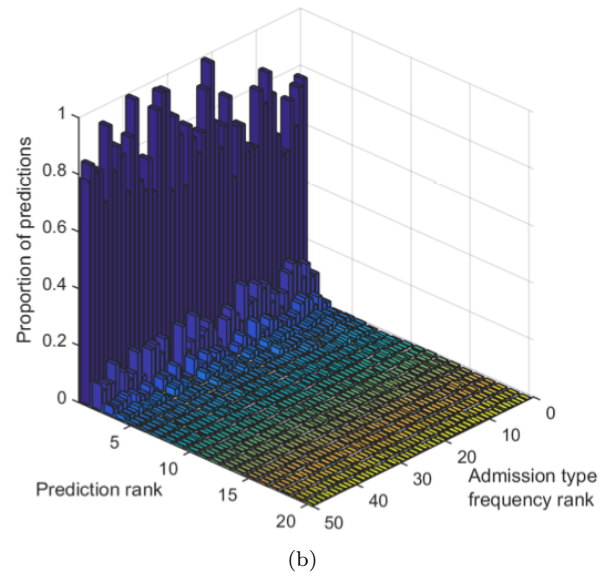
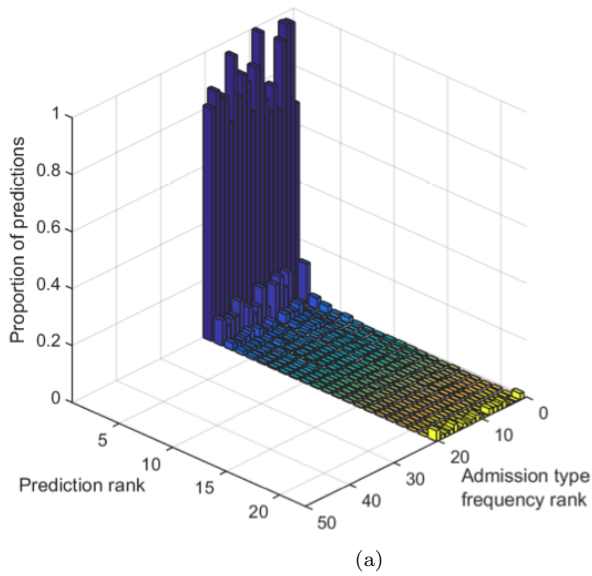


Figure 4: Prediction rank histograms across different DRG codes using (a) 20 vs. (b) 50 salient admission types.

records to predict future hospital admissions of a patient based on the patient’s previous medical history. In particular we scrutinized the original authors’ expectation that their preliminary results would scale as the number of explicitly modelled hospital admission types included in their history vector model is increased. Our experiments conducted on a real-world data corpus of nearly 3,000,000 admissions supports the authors’ claims and demonstrate that the high prediction accuracy is maintained well even when the number of admission types explicitly included in the model is increased to account for 98% of all admissions in the corpus.

Considering our findings, there are several well motivated directions for extending the adopted model. Firstly, the use of more granular DRG codes (i.e. codes corresponding to deeper hierarchical levels of the ICD’s disease classification tree) should be investigated. Secondly, the practical significance of the predictions would greatly benefit from the use of temporal information. Lastly, alternative criteria for the choice of salient admission types should be examined; for example, instead of selecting these based on their frequency in the data, the criterion could be based on the number of different individuals affected (thereby eliminating the skew effected by conditions which may be experienced by a small number of people but which are chronic in nature), their association with mortality, or their cost.

References

- [1] O. Arandjelović. Object matching using boundary descriptors. *In Proc. British Machine Vision Conference*, 2012. DOI: 10.5244/C.26.85.
- [2] O. Arandjelović. Reading ancient coins: automatically identifying denarii using obverse legend seeded retrieval. *In Proc. European Conference on Computer Vision*, 4:317–330, 2012.
- [3] O. Arandjelović. Discovering hospital admission patterns using models learnt from electronic hospital records. *Bioinformatics*, 31(24):3970–3976, 2015.
- [4] O. Arandjelović. Prediction of health outcomes using big (health) data. *In Proc. International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2543–2546, 2015.
- [5] O. Arandjelović. On the discovery of hospital admission patterns – a clarification. *Bioinformatics*, 2016.
- [6] A. Beykikhoshk, O. Arandjelović, D. Phung, and S. Venkatesh. Hierarchical Dirichlet process for tracking complex topical structure evolution and its application to autism research literature. *In Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 1:550–562, 2015.
- [7] A. Beykikhoshk, O. Arandjelović, D. Phung, S. Venkatesh, and T. Caelli. Using Twitter to learn about the autism community. *Social Network Analysis and Mining*, 5(1):5–22, 2015.
- [8] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior. The relation between the ROC curve and the CMC. *In Proc.*

- IEEE Workshop on Automatic Identification Advanced Technologies*, pages 15–20, 2005.
- [9] A. Bottle, P. Aylin, and A. Majeed. Identifying patients at high risk of emergency hospital admissions: a logistic regression analysis. *J R Soc Med*, 99(8):406–414, 2006.
 - [10] J. Butler and A. Kalogeropoulos. Hospital strategies to reduce heart failure readmissions. *J Am Coll Cardiol*, 60(7):615–617, 2012.
 - [11] A. De Gaetano, T. Hardy, B. Beck, E. Abu-Raddad, P. Palumbo, J. Bue-Valleskey, and N. Pørksen. Mathematical models of diabetes progression. *Am J Physiol Endocrinol Metab*, 295:E1462–E1479, 2008.
 - [12] K. Dharmarajan, A. F. Hsieh, Z. Lin, H. Bueno, J. S. Ross, I. Horwitz, J. A. Barreto-Filho, N. Kim, S. M. Bernheim, L. G. Suter, E. E. Drye, and H. M. Krumholz. Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia. *JAMA*, 309(4):355–363, 2013.
 - [13] B. Friedman, H. J. Jiang, and A. Elixhauser. Costly hospital readmissions and complex chronic illness. *Inquiry*, 45(4):408–421, 2008–2009.
 - [14] D. H. Gorski and S. P. Novella. Clinical trials of integrative medicine: testing whether magic works? *Trends Mol Med*, 20(9):473–476, 2014.
 - [15] C. H. Jackson, L. D. Sharples, S. G. Thompson, S. W. Duffy, and E. Couto. Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society, Series D*, 52(2):193–209, 2003.
 - [16] D. Kansagara, H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman, and S. Kripalani. Risk prediction models for hospital readmission: a systematic review. *JAMA*, 306(15):1688–1698, 2011.
 - [17] A. N. Long and S. Dagogo-Jack. Comorbidities of diabetes and hypertension: mechanisms and approach to target organ protection. *J Clin Hypertens*, 13(4):244–251, 1956.
 - [18] A. M. Mudge, K. Kasper, A. Clair, H. Redfern, J. J. Bell, M. A. Barras, G. Dip, and N. A. Pachana. Recurrent readmissions in medical patients: a prospective study. *J Hosp Med*, 6(2):61–67, 2011.
 - [19] W. Rieutort-Louis and O. Arandjelović. Bo(V)W models for object recognition from video. In *Proc. International Conference on Systems, Signals and Image Processing*, pages 89–92, 2015.
 - [20] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw Hill, New York, 1983.
 - [21] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. IEEE International Conference on Computer Vision*, 2:1470–1477, 2003.
 - [22] R. Sukkar, E. Katz, Y. Zhang, D. Raunig, and B. T. Wyman. Disease progression modeling using hidden Markov models. In *Proc. IEEE International Conference on Engineering in Medicine and Biology Society*, pages 2845–2848, 2012.
 - [23] B. Topp, K. Promislow, G. de Vries, R. M. Miura, and D. T. Finegood. A model of β -cell mass, insulin, and glucose kinetics: Pathways to diabetes. *J Theor Biol*, 206:605–619, 2000.
 - [24] World Health Organization. *International statistical classification of diseases and related health problems.*, volume 1. World Health Organization, 2004.
 - [25] W. Ye, D. J. M. Isaman, and J. Barhak. Use of secondary data to estimate instantaneous model parameters of diabetic heart disease: Lemonade Method. *Information Fusion*, 13:137–145, 2012.